Automatic Relevance Determination (ARD) with Statistical Guarantees

O'Bayes 2022, Santa Cruz

Zach Liu, Feng Liang

09/07/2022

Department of Statistics
University of Illinois at Urbana-Champaign

Introduction

Sparse High Dimensional Models

- Examples: regression, classification, graphical model, and network analysis
- To reliably learn a high-dimensional model based on a finite sample, we have to impose some structural assumptions.
- Sparsity: only a small fraction of the model parameters $\Theta = (\theta_1, \cdots, \theta_p)$ are non-zero.

Main techniques:

- Penalization framework
- Bayesian framework

Penalization Framework

The Penalized Likelihood Framework has the following form:

$$\underbrace{\hat{\Theta}}_{\text{Estimate}} \in \arg\min_{\Theta \in \Omega} \Big\{ \underbrace{-\log p(\text{Data} \mid \Theta)}_{\text{Loss function}} + \lambda \underbrace{\text{Pen}(\Theta)}_{\text{Penalty function}} \Big\}$$

For example, for linear regression, we have

$$\min_{\beta \in \mathbb{R}^p} \left[\|y - X\beta\|^2 + \lambda \sum_{j=1}^{p} \text{Pen}(\beta_j) \right]$$

where $\lambda \geq 0$ is a tuning parameter, and $\text{Pen}(\theta_j)$ is a sparsity-inducing penalty function, such as ℓ_0/ℓ_1.

Bayesian Framework

In the Bayesian framework, we have a generative model for both data and parameters:

$$\begin{aligned} \text{Prior} &: \pi(\Theta) \\ \text{Likelihood} &: P(\text{Data} \mid \Theta) \\ &\Longrightarrow \\ \text{Posterior} &: \pi(\Theta \mid \text{Data}) \end{aligned}$$

In fact, the prior $\pi(\Theta)$ plays the same role as a penalty function:

$$\text{Penalty} = \text{minus-log Prior}$$

Penalization, a Special Case of the Bayesian Framework

- The MAP estimate of Θ is the value that maximizes $\pi(\Theta \mid \text{Data})$. Recall

$$\pi(\Theta \mid \text{Data}) = \frac{P(\text{Data} \mid \Theta) \times \pi(\Theta)}{\int P(\text{Data} \mid \tilde{\Theta}) \times \pi(\tilde{\Theta}) d\tilde{\Theta}} \propto P(\text{Data} \mid \Theta) \times \pi(\Theta)$$

- So finding MAP is equivalent to minimizing

$$-\log P(\text{Data} \mid \Theta) + \underbrace{\big[-\log \pi(\Theta)\big]}_{\text{Bayesian Penalty}},$$

  that is, Penalty = minus-log Prior.

- For example, Lasso with penalty $\lambda|\theta| \Longrightarrow$ MAP of Double Exponential Prior.

Choice of Priors

Sparsity-inducing priors used in the Bayesian approach can be broadly classified into two categories.

- Unimodal continuous dist, such as
  - Double Exponential prior [Park and Casella, 2008]
  - Horseshoe prior [Carvalho et al., 2009]
  - global-and-local shrinkage prior [Polson and Scott, 2010]
- Two-group mixture dist, such as
  - spike-and-slab Normal prior [George and McCulloch, 1993; Ročková and George, 2014]
  - spike-and-slab Lasso prior [Ročková and George, 2016]

Normal Priors

Normal priors are usually **not recommended**.

- Not a heavy tail dist, i.e., $|\frac{\partial}{\partial\theta}\log\pi(\theta)|$ is not bounded [Johnstone and Silverman, 2004]
- Shrinkage but no sparsity. For example, ridge regression does not lead to a sparse coefficient vector.

However, what if the prior variance for each parameter can be set adaptively?

$$\pi(\theta_j) \sim N(0, r_j^2)$$

- r_j^2 = importance/relevance of θ_j
- In particular, $r_j^2 = 0$ for irrelevant parameters.

Normal Priors

Normal priors are usually **not recommended**.

- Not a heavy tail dist, i.e., $|\frac{\partial}{\partial\theta}\log\pi(\theta)|$ is not bounded [Johnstone and Silverman, 2004]
- Shrinkage but no sparsity. For example, ridge regression does not lead to a sparse coefficient vector.

However, what if the prior variance for each parameter can be set adaptively?

$$\pi(\theta_j) \sim N(0, r_j^2)$$

- r_j^2 = importance/relevance of θ_j
- In particular, $r_j^2 = 0$ for irrelevant parameters.

Automatic Relevance Determination (ARD)

MacKay (1995)

The ARD model puts a prior over the regression parameters which embodies the concept of relevance. This is done in a simple and soft way by introducing multiple regularisation constants, one associated with each input. Using Bayesian methods the regularisation constants for junk inputs are automatically inferred to be largely preventing those inputs from causing significant over-fitting.

- ARD Prior: $p(\Theta|\boldsymbol{r}^2) = \Pi_{j=1}^{p} \mathsf{N}(\theta_j|0, r_j^2)$
- $r_j^2 = 0$ implies a point mass posterior distribution at zero on θ_j
- Learn $\boldsymbol{r}^2$ by optimizing evidence function

Prior work

[MacKay, 1995; Neal, 1995]: ARD for single layer neural network

[Tipping, 2001; Tipping and Faul, 2003]: ARD algorithm for relevance vector machine

[Wipf and Nagarajan, 2007]: Alternative view and optimization method for ARD

[Titsias and Lázaro0Gredilla, 2014; Kharitonov et al., 2018]: Variational ARD Algorithm for Bayesian Neural Network

Main focus has been on developing algorithms for prediction, no statistical guarantees or theoretical results for estimation and variable selection.

Outline

1. Introduction
2. Method
3. Theoretical Results
4. Simulation

Method

Variational ARD

Linear Regression Model:

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \mathsf{N}(\mathbf{0}, \sigma^2 I_n)$$

ARD Prior:

$$p(\boldsymbol{\beta}) = \Pi_{j=1}^{p} p_j(\beta_j) = \Pi_{j=1}^{p} \mathsf{N}(\beta_j \mid 0, r_j^2)$$

Variational Distribution:

$$q(\boldsymbol{\beta}) = \Pi_{j=1}^{p} q_j(\beta_j) = \Pi_{j=1}^{p} \mathsf{N}(\beta_j \mid \mu_j, \phi_j^2)$$

Scale-Invariance Property

Effect of scale transformation:

$$\underbrace{\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\phi}}^2, \hat{\boldsymbol{r}}^2, \hat{\sigma}^2}_{\text{original optimal solution}} \implies \text{multiply } \boldsymbol{x}_j \text{ by } c \implies \underbrace{\begin{cases} \text{Divide } \hat{\mu}_j \text{ by } c \\ \text{Divide } \hat{\phi}_j^2 \text{ by } c^2 \\ \text{Divide } \hat{r}_j^2 \text{ by } c^2 \end{cases}}_{\text{transformed optimal solution}}$$

Scale transformation to any feature $\boldsymbol{x}_j$ won't change the outcome of estimation and variable selection

Remark:

1. Subset selection or equivalently ℓ_0 penalty is scale-invariant.
2. Lasso, Ridge, and some Bayesian methods are not scale-invariant

Evidence Lower Bound Objective

Minimize ELBO

$$\begin{aligned}\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\phi}, \boldsymbol{r}, \sigma^2) &= -\,\mathsf{E}_q \log p(\boldsymbol{y}|\boldsymbol{\beta}) + \alpha \cdot \mathsf{KL}(q\|p) \\ &= -\,\mathsf{E}_q \log \mathsf{N}(\boldsymbol{y}|\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 I_n) + \alpha \cdot \sum_{j=1}^{p} \mathsf{KL}\left(\mathsf{N}(\mu_j, \phi_j^2) \| \, \mathsf{N}(0, r_j^2)\right)\end{aligned}$$

where

- $\boldsymbol{\alpha} \geq 0$ is a hyperparameter [Higgins et al., 2017; Yang et al., 2020]
- E_q = expectation w.r.t the variational dist $\beta_j \sim \mathsf{N}(\mu_j, \phi_j^2)$

$$-\,\mathsf{E}_q \log p(\boldsymbol{y}|\boldsymbol{\beta}) = \frac{n}{2} \log \sigma^2 + \frac{\|\boldsymbol{y} - \mathbf{X}\boldsymbol{\mu}\|^2}{2\sigma^2} + \sum_j \frac{\phi_j^2}{2\sigma^2} \|\boldsymbol{x}_j\|^2$$

$$\mathrm{KL}\left(\mathsf{N}(\mu_j, \phi_j^2) \middle|\middle| \mathsf{N}(0, r_j^2)\right) = \begin{cases} \frac{1}{2}\left(\log r_j^2 - \log \phi_j^2 + \frac{\mu_j^2+\phi_j^2}{r_j^2} - 1\right), & \phi_j^2 \neq 0, r_j^2 \neq 0 \\ 0, & \mu_j = \phi_j^2 = r_j^2 = 0 \\ +\infty & \text{otherwise} \end{cases}$$

Note that the prior variances $\boldsymbol{r} = (r_1, \ldots, r_p)$ only appear in the expression above. It is easy to verify that the term above (or equivalently $\mathcal{L}$) is minimized at

Pythagorean

$$r_j^2 = \mu_j^2 + \phi_j^2, \quad j = 1, \ldots, p.$$

Computation: Coordinate Descent

Using the Pythagorean relationship:

$$r_j^2 = \mu_j^2 + \phi_j^2,$$

we can eliminate $\boldsymbol{r}^2$ and reduce the ELBO to

$$\left\{ \frac{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\mu}\|_2^2}{2n} + \sum_{j=1}^{p} \left[\frac{\phi_j^2}{2n} \|\boldsymbol{x}_j\|_2^2 + \frac{\sigma^2 \alpha}{n} \cdot \rho(\mu_j, \phi_j^2) \right] \right\} \cdot \frac{n}{\sigma^2} + \frac{n}{2} \log \sigma^2;$$

then derive a Coordinate Descent algorithm: optimize w.r.t $(\mu_j, \phi_j^2), j = 1, \cdots, p$ and σ^2 sequentially while fixing other parameters.

Algorithm 1: Coordinate Descent

Input $(\boldsymbol{X}, \boldsymbol{y}, \alpha)$;

Init $\hat{\boldsymbol{\mu}}$ and $\hat{\sigma}^2$;

while *Not Converge* **do**

  for j *in* $1, \ldots, p$ **do**

    $\hat{\boldsymbol{z}}_j = \boldsymbol{y} - \sum_{k \neq j} \boldsymbol{x}_k \hat{\mu}_k$;

    $\hat{\mu}_j = \frac{\boldsymbol{x}_j^T \hat{\boldsymbol{z}}_j}{\|\boldsymbol{x}_j\|_2^2} \left(1 - \frac{\alpha \hat{\sigma}^2 \|\boldsymbol{x}_j\|_2^2}{(\boldsymbol{x}_j^T \hat{\boldsymbol{z}}_j)^2}\right)_+$;

    $\hat{\phi}_j^2 = \frac{\alpha \hat{\sigma}^2}{\|\boldsymbol{x}_j\|_2^2} \left(1 - \frac{\alpha \hat{\sigma}^2 \|\boldsymbol{x}_j\|_2^2}{(\boldsymbol{x}_j^T \hat{\boldsymbol{z}}_j)^2}\right)_+$;

  $\hat{\sigma}^2 = \frac{\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\mu}}\|_2^2 + \sum_{j=1}^p \hat{\phi}_j^2 \|\boldsymbol{x}_j\|_2^2}{n}$;

$\hat{\boldsymbol{r}}^2 = \hat{\boldsymbol{\mu}}^2 + \hat{\boldsymbol{\phi}}^2$;

- Note that $\alpha\hat{\sigma}^2$ always show up together

Algorithm 2: Coordinate Descent (Alternative)

Input $(\boldsymbol{X}, \boldsymbol{y}, \alpha)$;

Init $\hat{\boldsymbol{\mu}}$;

while *Not Converge* **do**

  for j *in* $1, \ldots, p$ **do**

    $\hat{\boldsymbol{z}}_j = \boldsymbol{y} - \sum_{k \neq j} \boldsymbol{x}_k \hat{\mu}_k$;

    $\hat{\mu}_j = \frac{\boldsymbol{x}_j^T \hat{\boldsymbol{z}}_j}{\|\boldsymbol{x}_j\|_2^2} \left(1 - \frac{\tilde{\alpha}\|\boldsymbol{x}_j\|_2^2}{(\boldsymbol{x}_j^T \hat{\boldsymbol{z}}_j)^2}\right)_+$;

    $\hat{\phi}_j^2 = \frac{\tilde{\alpha}}{\|\boldsymbol{x}_j\|_2^2} \left(1 - \frac{\tilde{\alpha}\|\boldsymbol{x}_j\|_2^2}{(\boldsymbol{x}_j^T \hat{\boldsymbol{z}}_j)^2}\right)_+$;

$\hat{\boldsymbol{r}}^2 = \hat{\boldsymbol{\mu}}^2 + \hat{\boldsymbol{\phi}}^2$;

- pre-specify a sequence of $\tilde{\alpha}$
- cross-validation on each $\tilde{\alpha}$
- Compute $\hat{\sigma}^2$ at the end, if needed.

Theoretical Results

Setup

Assume $\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta}^* + \mathrm{N}(\mathbf{0}, \sigma^2 I_n)$ and σ^2 known.

$$\text{Truth} \quad : \quad \beta^*, \quad S = supp(\boldsymbol{\beta}^*), \quad s = |S|.$$

Output a variational dist, not posterior dist:

$$\hat{q}(\boldsymbol{\beta}) = \prod_j N(\beta_j | \hat{\mu}_j, \hat{\phi}_j^2)$$

- Estimation Consistency: $\|\hat{\boldsymbol{\mu}} - \boldsymbol{\beta}^*\|_1, \|\hat{\boldsymbol{\mu}} - \boldsymbol{\beta}^*\|_2, \|\hat{\boldsymbol{\mu}} - \boldsymbol{\beta}^*\|_\infty$
- Selection Consistency: $sign(\hat{\boldsymbol{\mu}}) = sign(\boldsymbol{\beta}^*)$, $supp(\hat{\boldsymbol{\phi}}^2) = supp(\boldsymbol{\beta}^*)$
- Variational Concentration Result for $\hat{q}(\boldsymbol{\beta})$.

A View of Penalized Regression

(1) **Pythagorean:** $r_j^2 = \mu_j^2 + \phi_j^2$

(2) **Optimal Variational Variance:** $\phi_j^2 = \tau(\mu_j)$

(1) + (2) $\Longrightarrow$ ELBO reduces to (assume σ^2 known):

$$L(\boldsymbol{\mu}) = \frac{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\mu}\|_2^2}{2n} + \sum_{j=1}^{p} g(\mu_j)$$

where g plays the role of a penalty function,

$$g(\mu) = \frac{\sigma^2\alpha}{2n}\Big[\frac{n\tau(\mu)}{\sigma^2\alpha} - \log(1 - \frac{n\tau(\mu)}{\sigma^2\alpha})\Big]$$

(Note: we normalized each column of $\boldsymbol{X}$ to have mean zero and norm $\sqrt{n}$.)

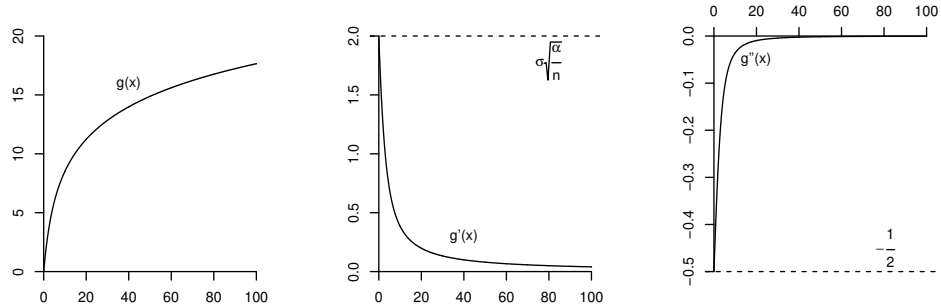

Figure 1: Penalty $g(x)$ (**left**); its derivative (**middle**); its second order derivative (**right**).

Solution Set

Reduced Objective:

$$L(\boldsymbol{\mu}) = \frac{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\mu}\|_2^2}{2n} + \sum_{j=1}^{p} g(\mu_j)$$

We analyze the following solutions:

- $\hat{\boldsymbol{\mu}}$: Any local optima within an ℓ_1 ball [Loh and Wainwright, 2015; Gan et al., 2020]

  $$\hat{\boldsymbol{\mu}} \in \{\|\boldsymbol{\mu}\|_1 \leq R\}$$

  Note: R can increase to infinity with sample size n.

- $\hat{\boldsymbol{\phi}}^2 = \boldsymbol{\tau}(\hat{\boldsymbol{\mu}})$, $\hat{q}(\boldsymbol{\beta}) = \prod_{j=1}^{p} \hat{q}_j(\beta_j) = \prod_{j=1}^{p} \mathsf{N}(\beta_j | \hat{\mu}_j, \hat{\phi}_j^2)$

Restricted Strong Convexity (RSC)

Assumption 1 (The RSC Condition)

There exists strictly positive constant C_1, C_2, for any vector $\Delta \in \mathbb{R}^p$,

$$\frac{\|\boldsymbol{X}\Delta\|_2^2}{n} \geq C_1\|\Delta\|_2^2 - C_2\frac{\log p}{n}\|\Delta\|_1^2 \tag{1}$$

- A caveat in high dimensions: $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\mu}\|_2^2$ is not strongly convex
- RSC is weaker than strong convexity, and
- can hold with high prob if X are sampled from Gaussian [Raskutti et al., 2010; Raskutti et al., 2011]

ℓ_1/ℓ_2 Error Bounds

Theorem 1

Suppose Assumption 1 holds, and

$$\|\boldsymbol{\beta}^*\|_1 \leq R < O\left(\sqrt{\frac{n}{\log p}}\right), \quad \alpha \asymp \log p,$$

then with high probability,

$$\|\hat{\boldsymbol{\mu}} - \boldsymbol{\beta}^*\|_2 \precsim \sqrt{\frac{s \log p}{n}}$$

$$\|\hat{\boldsymbol{\mu}} - \boldsymbol{\beta}^*\|_1 \precsim s\sqrt{\frac{\log p}{n}}$$

$$\|\hat{\boldsymbol{\phi}}^2\|_1 \precsim \frac{s \log p}{n}$$

Conditions on Correlations Among Features

Assumption 2

There exist positive constants c_∞ *and* η *such that,*

$$\|(\frac{\boldsymbol{X}_S^T \boldsymbol{X}_S}{n})^{-1}\|_\infty \leq c_\infty \tag{2}$$

and

$$\|\boldsymbol{X}_{S^c}^T \boldsymbol{X}_S (\boldsymbol{X}_S^T \boldsymbol{X}_S)^{-1}\|_\infty < \eta \tag{3}$$

- Condition (2) is common
- Condition (3) is weaker than the irrepresentable condition in [Wainwright, 2009; Loh and Wainwright, 2017] where η needs to be strictly less than 1, while our η can be any positive constant.

ℓ_∞ Error Bounds and Selection Consistency

Theorem 2

Suppose Assumptions 1 and 2 hold,

$$\|\boldsymbol{\beta}^*\|_1 \leq R < O\left(\sqrt{\frac{n}{\log p}}\right), \quad \alpha \asymp \log p,$$

$$|\boldsymbol{\beta}^*_S|_{\min} \geq O\left(\sqrt{\frac{\log p}{n}}\right), \quad n \gtrsim s \log p,$$

then with high probability,

1. $\hat{\boldsymbol{\mu}}$ *is unique with* $\|\hat{\boldsymbol{\mu}} - \boldsymbol{\beta}^*\|_\infty \precsim \sqrt{\frac{\log p}{n}}$
2. $sign(\hat{\boldsymbol{\mu}}) = sign(\boldsymbol{\beta}^*)$
3. $supp(\hat{\boldsymbol{\mu}}) = supp(\hat{\boldsymbol{\phi}}^2) = supp(\boldsymbol{\beta}^*) = S$

Concentration

- Small regions around the truth with radius $\xi_n(\to 0)$, e.g.,

$$\mathcal{B}_n = \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_1 \leq \xi_n\}, \text{ or } \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \leq \xi_n \text{ or } \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_\infty \leq \xi_n\}$$

- Martin [2021] proposed to study $\mathsf{E}_{\boldsymbol{\beta}^*}(\hat{\Pi}(\mathcal{B}_n))$ when $\hat{\Pi}$ is a data-dependent measure (not necessarily posterior) over the parameter space
- We study whether

$$\mathsf{E}_{\boldsymbol{\beta}^*}(\hat{q}(\mathcal{B}_n)) \to 1$$

where

$$\hat{q}(\boldsymbol{\beta}) = \Pi_{j=1}^p \hat{q}_j(\beta_j) = \Pi_{j=1}^p \mathsf{N}(\beta_j | \hat{\mu}_j, \hat{\phi}_j^2)$$

is our variational dist

Concentration

- Small regions around the truth with radius $\xi_n(\to 0)$, e.g.,

$$\mathcal{B}_n = \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_1 \leq \xi_n\}, \text{ or } \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \leq \xi_n \text{ or } \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_\infty \leq \xi_n\}$$

- Martin [2021] proposed to study $\mathsf{E}_{\boldsymbol{\beta}^*}(\hat{\Pi}(\mathcal{B}_n))$ when $\hat{\Pi}$ is a data-dependent measure (not necessarily posterior) over the parameter space
- We study whether

$$\mathsf{E}_{\boldsymbol{\beta}^*}(\hat{q}(\mathcal{B}_n)) \to 1$$

where

$$\hat{q}(\boldsymbol{\beta}) = \Pi_{j=1}^p \hat{q}_j(\beta_j) = \Pi_{j=1}^p \mathsf{N}(\beta_j | \hat{\mu}_j, \hat{\phi}_j^2)$$

is our variational dist

Variational Concentrates on the Truth?

Theorem 3

Define

$$\mathcal{B}_n = \{\boldsymbol{\beta} : supp(\boldsymbol{\beta}) = S, \|\boldsymbol{\beta}_S - \boldsymbol{\beta}^*_S\|_\infty \leq \xi_n\}$$

Assume conditions of Theorem 2 hold. Then as $n \to +\infty$*,*

$$E_{\boldsymbol{\beta}^*}(\hat{q}(\mathcal{B}_n)) \to 1$$

holds for any series $\xi_n \asymp \left(\sqrt{\frac{\log p}{n}}\right)^{1-\delta}$*, where* δ *is a constant that can take arbitrary small value in* $(0, 1)$*.*

Theorem 3 implies that:

- $\mathsf{E}_{\boldsymbol{\beta}^*}(\hat{q}\{\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_1 \leq s \cdot \xi_n\}) \to 1$
- $\mathsf{E}_{\boldsymbol{\beta}^*}(\hat{q}\{\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|_2 \leq \sqrt{s} \cdot \xi_n\}) \to 1.$

Simulation

We compared our algorithm with the following Bayesian selection algorithms,

- sparsevb [Ray and Szabó, 2021]
- varbvs [Carbonetto and Stephens, 2012]
- ebreg [Martin et al., 2017]
- EMVS [Ročková and George, 2014]

on simulated data using Lasso (glmnet) [Friedman et al., 2010] as a benchmark.

Experiment 1

Each element of $\boldsymbol{X}$ are i.i.d N$(0, 1)$:

- Case 1: $(n, p, s, \sigma) = (200, 800, 10, 1), \boldsymbol{\beta}_S^* = (1, -2, 3, -4, \cdots, 9, -10)$ (none zero coefficient at random place);
- Case 2: $(n, p, s, \sigma) = (200, 1000, 15, 1), \boldsymbol{\beta}_S^* = 10$ (at random place);
- Case 3: $(n, p, s, \sigma) = (100, 400, 20, 5), \boldsymbol{\beta}_S^* = \log 100$ (at the beginning);
- Case 4: $(n, p, s, \sigma) = (100, 400, 20, 5), \boldsymbol{\beta}_S^* = 2 \log 100$ (at the end).

Table 1: Uncorrelated Design

Metric	Method	Case 1	Case 2	Case 3	Case 4
ℓ_2 error	lasso	0.719 ± 0.093	1.13±0.106	12.4±3.22	16.4±8.74
	sparsevb	0.264±0.075	**0.293±0.064**	8.12±6.02	**7.35±6.52**
	varbvs	**0.223±0.050**	0.293±0.065	14.1±6.66	24.0±16.9
	ebreg	0.236±0.051	0.308±0.063	11.3±6.9	8.35±12.6
	emvs	0.608±0.044	0.577±0.049	11.8±6.57	11.9±14.3
	ours	0.235±0.049	0.296±0.066	**5.82±4.76**	8.07±7.78
FDR	lasso	0.541±0.142	0.405±0.133	0.603±0.09	0.638±0.063
	sparsevb	0.033±0.059	**0±0**	0.119±0.177	0.062±0.108
	varbvs	0.001±0.009	0.001±0.009	**0.076±0.122**	0.064±0.128
	ebreg	**0±0**	**0±0**	0.146±0.203	0.065±0.182
	emvs	0.005±0.024	**0±0**	0.251±0.162	0.199±0.225
	ours	0.036±0.052	**0±0**	0.324±0.136	**0.042±0.154**
TPR	lasso	1±0	1±0	0.891±0.173	0.956±0.117
	sparsevb	1±0	1±0	0.844±0.252	**0.975±0.108**
	varbvs	1±0	1±0	0.409±0.369	0.461±0.435
	ebreg	1±0	1±0	0.669±0.323	0.905±0.236
	emvs	1±0	1±0	0.698±0.3	0.893±0.22
	ours	1±0	1±0	**0.948±0.127**	0.967±0.124

Experiment 2

Each row of $\boldsymbol{X}$ are i.i.d N$(\mathbf{0}, \Sigma)$, where diagonal elements of Σ are 1 and off diagonal elements of Σ are ς. For all three cases, $(n, p, s, \sigma) = (200, 400, 40, 1)$ and none zero coefficient index is randomly generated.

- Case 1: ς = 0.2
- Case 2: ς = 0.5
- Case 3: ς = 0.8

Table 2: Correlated Design

Metric	Method	$\varsigma = 0.2$	$\varsigma = 0.5$	$\varsigma = 0.8$
ℓ_2 error	lasso	4.59+0.719	7.53+1.38	10.8+1.96
	sparsevb	0.896±0.217	2.96±0.546	7.7±1.67
	varbvs	**0.553±0.069**	**0.707±0.098**	51.0±9.2
	ebreg	1.71±0.31	4.8±0.723	3.94±4.7
	emvs	1.22±0.124	1.54±0.182	**2.14±0.207**
	ours	0.561±0.069	0.797±0.141	2.18±0.478
FDR	lasso	0.54±0.044	0.573±0.037	0.605±0.037
	sparsevb	**0±0**	0.003±0.009	0.014±0.017
	varbvs	0.001±0.005	0.002±0.006	0.453±0.134
	ebreg	**0±0**	**0±0**	0.008±0.026
	emvs	0.040±0.032	0.058±0.05	0.041±0.034
	ours	0.003±0.009	**0±0**	**0±0**
TPR	lasso	0.995±0.01	0.979±0.023	0.956±0.029
	sparsevb	0.993±0.011	0.913±0.019	0.768±0.033
	varbvs	**1±0**	**1±0**	0.214±0.083
	ebreg	0.949±0.0145	0.828±0.027	0.922±0.156
	emvs	**1±0**	**1±0**	**1±0**
	ours	**1±0**	1±0.002	0.953±0.021

Conclusions

- We study the Automatic Relevance Determination (ARD) model for high-dimensional linear regression under sparsity constraints.
- ARD introduces an individual relevance/variance parameter for each regression coefficient, which we propose to learn via variational optimization.
- When relevance/variance is set to zero, corresponding features are automatically filtered out.
- For our variational solutions, we establish convergence results, in terms of parameter estimation and variable selection, which provide a theoretical justification for ARD models.